

A Survey: Evaluation of Ensemble Classifiers and Data Level Methods to Deal with Imbalanced Data Problem in Protein-Protein Interactions

Seena Mary Augusty^{*1}, Sminu Izudheen²

Department of Computer Science, Rajagiri School of Engineering and Technology, India

^{*1}seenamaryaugusty@gmail.com; ²sminu_i@rajagiritech.ac.in

Abstract

Over the past few decades, protein interactions have gained importance in many applications of prediction and data mining. They aid in cancer prediction and various other disease diagnosis. Imbalanced data problem in protein interactions can be resolved both at data as well as algorithmic levels. This paper evaluates and surveys various methods applicable at data level as well as ensemble methods at algorithmic level. Cluster based under sampling, over sampling along with data based methods were evaluated under Data level. Ensemble classifiers were evaluated at the algorithmic level. Unstable base classifiers such as SVM and ANN can be employed for ensemble classifiers such as Bagging, Adaboost, Decorate, Ensemble non-negative matrix factorization and so on. Random forest can improve the ensemble classification in dealing with imbalanced data problem over Bagging as well as Adaboost method for high dimensional data.

Keywords

Bagging; Adaboost; Decorate; Oversampling; Under sampling

Introduction

Protein-protein interactions play numerous important roles in many cellular functions, including DNA replication, transcription and translation, signal transduction, and metabolic pathways. Thereby, aiding in diagnosis and prognosis of various diseases. Recently, a huge increase in the number of protein-protein interactions has made the prediction of unknown protein-protein interactions important for the understanding of living cells. However, the protein-protein interactions experimentally obtained so far are often incomplete and contradictory and, consequently, computational methods now have an upper hand in predictions. These prediction methods have been reviewed for classification under which

each has its own advantage over the other. The significant difficulty and frequent occurrence of the class imbalance problem indicate the need for extra research efforts. This paper extensively evaluates recent developments in the field of solving imbalanced data problem and subsequently classifying the new solutions under each category. Finally proposing a slight enhancement for the solution of integrated cluster under sampling with ensemble classifiers, replacing bagging and Adaboost with Random forest for the paper (yongqing et al 2012). The combining method employed is Majority voting of all the decision trees.

Reviews under each Category

The insight gained from the comprehensive analysis of various solutions for handling imbalanced data problem, are reviewed in this paper.

Imbalanced Data Problem in Various Disciplines

Imbalanced data problem arises when the number of interacting pairs is very much less than the number of non interacting pairs. Former is known as positive dataset and the latter is known as negative samples. Protein in the same sub cellular location is seen as positive sample and in non sub cellular location is seen as negative sample. Various methods for wide range of applications to solve imbalanced data problem are present which can be used to check their compatibility with the protein interaction domain. One such generalisation of binary cases is described in paper (Victoria López et al 2012). This focuses the intrinsic behaviours of the imbalanced data problem such as class overlap and dataset shift. It is a cost sensitive learning solution that integrates model at both data as well as algorithmic level under the assumption that

higher misclassification occur at the minority samples and is sought after minimisation high cost errors.

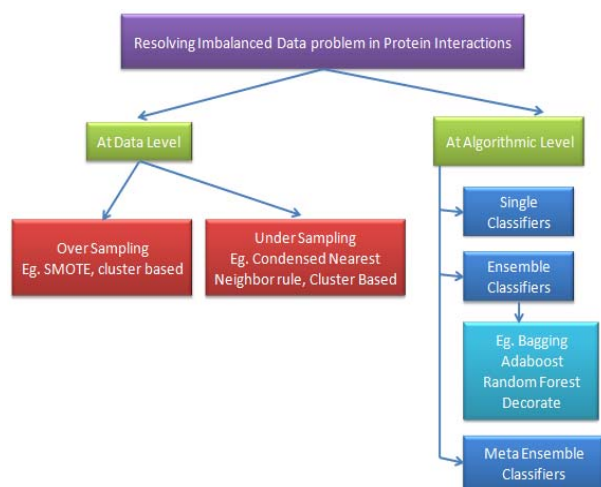


FIG. 1 CLASSIFICATION HIERARCHY

The imbalanced data problem is relaxed in unsupervised self organising learning with support vector ranking as mentioned in (Yok-Yen Nguwi et al 2010). In this method variables are selected by the model adopted by support vector machines to deal with this problem. ESOM also known as Emergent Self-Organising Map is used to cluster the ranker features so as to provide for unsupervised cluster classification. A Kolomogrov-Smirnov statistic based on decision tree method(K-S tree) (Rongsheng Gong et al 2012) is a latest method in which complex problem is divided into several easier sub problems, in that case imbalanced distribution becomes less daunting. This method is also used for feature selection removing the redundant ones. After division, a two-way re-sampling is employed to determine optimal sampling criteria and rebalanced data is used to incorporate into logistic regression models. Thus distribution of the dataset is used as an advantage for this method. Recently, information granulation based data mining (Mu-Chen Chen et al 2008) has gained a wide acceptance which uses the concept of human ability to process the information tackles the imbalanced data problem. While balancing the accuracies over the classes it may result in increase of accuracy over minority class whereas the other decreases. So in the multi-objective optimisation approach for class imbalance learning (Paolo Soda et al 2011) achieves global accuracy by the choice driven by the parameter on the validation set and, between the output of a classifier trained on the original skewed distribution and the output of a classifier trained according to a learning method addressing the course of imbalanced data. Figure 1 shows the

classification under which each solution can be categorised.

At Data Level

Sampling is done at the data level in which either minority sample size is increased as in over sampling or majority sample size is reduced as in under sampling. Methods utilising these two techniques are reviewed under each category. Both under sampling and oversampling can be incorporated with an ensemble of SVM which can improve prediction as mentioned in the paper (Yang Liu et al 2011). Pre-processing is an important tool for dealing with uneven distribution of the dataset. The paper (Alberto Fernández et al 2009) revisits a new concept of adaptive inference system with parametric conjunction operators on the fuzzy rule based classification system.

1) Preprocessing

Other way of tackling the inequitable distribution of dataset is by pre-processing the data beforehand to the learning process. In this paper (Salvador Garcí'a et al 2012), an exemplar that accomplishes learning process by storing entities in the Euclidean n-space. Prediction of the incoming dataset is performed by computing their distance to the nearest exemplar. This exemplar is chosen based on the evolutionary algorithms. Analysis of an evolutionary RBFN design algorithm, CO²RBFN, a evolutionary cooperative-competitive model for imbalanced data sets (María Dolores Pérez-Godoy et al 2010) is made. It can work well with pre-processing method such as SMOTE. As in (Francisco Fernández-Navarro et al 2011), in the first stage, the minority class is applied with over sampling procedure to balance in part the size of the classes. Then, the MA(memetic algorithm) is run and the data are again over-sampled in different generations of the evolution, generating new patterns of the minimum sensitivity class. MA optimises radial basis functions neural network (RBFNNs). These methods include different over-sampling procedures in the pre-processing stage, a threshold-moving method where the output threshold is transversed toward inexpensive classes and ensembles approaches combining the models obtained with these techniques overcomes to a great extent of the imbalanced data problem. Pre-processing unbalanced data using SVM (M.A.H. Farquard et al 2012) first employs SVM as a pre-processor and then the target values of SVM

are replaced by the predictions of trained SVM in turn are used to train Multilayer Perceptron (MLP), Logistic Regression (LR), and Random Forest (RF). This method efficiently tackles the uneven distribution of dataset.

2) *Over Sampling*

Minority kind of sample is clustered using K-means and subsequently using genetic algorithm to gain a new sample which has valid information as proposed in (Yang Yong et al 2012) could be employed to enhance the performance of the minority kind in the imbalanced data set. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems (Ming Gao et al 2011), is a powerful technique for integrating the synthetic minority over-sampling technique (SMOTE) and the particle swarm optimisation (PSO) and radial basis function (RBF) classifier. Synthetic instances for the positive class is generated by SMOTE in order to balance the training data set. Then RBF classifier is constructed based on the over sampled training data. Cluster based under sampling is demonstrated effective in (Show-Jane Yen et al 2009) solving imbalanced distribution by removing the clusters of the majority classes which are nearer to the minority class. Over sampling can be done by simple random sampling in which high variance produced by the Horvitz-Thompson estimator is used as the paramount characteristics for re sampling. In paper (Nicolás García-Pedrajas et al 2011), misclassified instances are used to find supervised projections and over sampling concepts are also defined.

3) *Under Sampling*

Cluster based under sampling is prominent in the paper (Show-Jane Yen et al 2009) which aims at resolving imbalanced data distribution. Training data selection needs to be taken care of well as the classifier can predict the incoming data belongs to majority class if most of the representative data are taken from the majority class. Here comes the relevance of under sampling in the imbalanced data distribution. The protein domain detection (Shu-xue Zou et al 2008) is first taken as an imbalanced data learning problem and this method is based on analyzing multiple sequence alignments. An under sampling method is put forward on distance-based maximal entropy in the feature space of SVM. Consequently, it helps in

predicting 3D structure of a protein as well as in the machine learning system on imbalanced datasets. Imbalanced data problem is dealt in (Der-Chiang Li et al 2010) by exploiting under sampling of dataset by mega-trend diffusion membership function for the minority class, and over sampling by building up the Gaussian type fuzzy membership function and α -cut to reduce the data size. It is found to be effective in solving unbalanced data by the usage of clustering based under sampling and then ensemble as discussed in (Pilsung Kang et al 2012). A novel approach of inverse random under sampling (IRUS) is proposed in (Muhammad Atif Tahir et al 2012). A composite decision boundary is constructed between majority class and minority class based on the training set produced by extensively under sampling the majority class. Promising results have been found out for this under sampling techniques outperforming all other classical under sampling techniques. Condensed nearest neighbour rule stores subset of the dataset which has efficient implementation of the nearest neighbour decision rule. Tomek has found yet another subset which makes the training set consistent known as Tomek links (in Gabriel graph). A new counterexample to Tomek's consistency theorem has been proved in (Godfried T Toussaint et al 1994). So this paves yet another path to solving data imbalanced problem at data level. Cost sensitive learning (Charles Elkan et al 2001) can be applied for optimal cost-sensitive classification which makes changes in the proportion of the negative sample generalised beneath under sampling technique.

At Algorithmic Level

Learning and building of models is accomplished in the algorithmic level. Either a single classifier or ensemble of classifiers can be employed. Algorithms are classified accordingly to the above mentioned criteria.

1) Single Classifiers and Computational Methods

Margin calibration in SVM class-imbalanced learning (Chan-Yun Yang et al 2009) utilises the identification of reference cost sensitive prototype as a penalty-regularized model. This method adopts an inversed proportional regularised penalty to re-weight the imbalanced classes. Then two regularisation factors such as penalty and margin is yielded to output unbiased classification.

Imbalanced learning tasks cannot be handled by conventional SVM as they tend to classify the entities of majority class which is a less important class. In order to solve this problem, a method known as Learning SVM with weighted margin criterion for classification of imbalanced data (Zhuangyuan Zhao et al 2011) is exploited. Here a weighted maximum margin criterion to optimize the data-dependent kernel is observed. Hence, giving chance to the minority class of being more clustered. The weight parameters are embedded in the Lagrangian SVM formulation is employed for imbalanced data classification problem via SVM with quadratic cost function (Jae Pil Hwang et al 2011). When protein dataset are stored in multi-relational database (Chien-I Lee et al 2008), multi-relational g-mean decision tree algorithm is used to solve imbalanced data problem. Multivariate statistical analyses is depicted to improve efficiency in classifiers (Hung-Yi Lin et al 2012). This multivariate statistical analyses solve problems which are stalled by high dimensionality hence improves classification training time. A novel approach of combining ANOVA (analysis of variance), FCM (Fuzzy clustering algorithm), and BFO (bacterial foraging optimisation) is put forward as new computational method for unbalanced data (Chou-Yuan Lee et al 2012), by first selection of beneficial feature subsets (by ANOVA), then clustering data into membership degrees (by FCM) and finally convergence is provided by yielding of global optima (by BFO). Two class learning for SVM (Raskutti B. Et al 2004) is investigated in which aggressive dimensionality reduction is done to improve the classification.

2) Ensemble Classifiers

In recent years there has been development in the field of ensemble classifiers in which the advantages of all single classifiers are combined together to yield a better prediction. Ensemble methods are widely used in various disciplines such as in (Larry Shoemaker et al 2008) where classifier ensembles is used to label spatially disjoint data. The combining method employed here is the probabilistic majority voting. Combination of ensemble learning with cost sensitive learning is proposed in different realm in (Jin Xiao et al 2012). These techniques can be utilised in protein interaction domain as it is dealt with imbalanced data problem. In this paper (Jin

Xiao et al 2012), combination of ensemble learning with cost sensitive learning yields a new version known as dynamic classifier ensemble method for imbalanced data (DCEID). Eventually new cost-sensitive selection criteria for Dynamic Classifier Selection (DCS) and Dynamic Ensemble Selection (DES) are constructed respectively to enhance the classification capability for imbalanced data. In pattern recognition realm, feature extraction is seen as imbalanced data problem for both negative and positive features. This method (Jinghua Wang et al 2012) can be generalised to all domains. This observation (Jinghua Wang et al 2012) covers two models in which first model relates to candidate extractors for minimising the other class and the latter one does vice versa. This combination is less likely to be affected by the imbalanced data problem. Ensemble methods by binarization technique focusing on one-vs-one and one-vs-all decomposition strategies proved to be efficient in (Mikel Galar et al 2011) for solving multi-class problems. Here empirical analysis of different aggregations is used to combine the outputs. In the neuro computing domain, model parameter selection via alternating SVM and gradient steps to minimise generalization error is employed (Todd W. Schiller et al 2010) which can be extended to protein interaction domain. Ensemble of SVM proved to be effective in this case. The protein sub cellular location is studied through CE-Ploc learning mechanism which is a ensemble approach combining the predictions of the base learners such as SVM, nearest neighbour, probabilistic neural network covariant discriminant produced prediction accuracy of about 81.47% using jack knife test. Classifier ensemble selection can be done using hybrid genetic algorithm (Young-Won Kim et al 2008). Ensemble can be constructed carefully emphasising the accuracies of the individual classifiers based on the use of supervised projections, both linear and non-linear (Nicolás García-Pedrajas et al 2011).

Meta Learning Regime

Protein structure classification is calculated by meta learners boosted and bagged meta learners but random forest outperformed all the other meta learners with the cross validated accuracy of 97.0%. Bagging and Adaboost can generally be adapted to its usage in vector quantization (Noritaka Shigei et al 2009). Bagging can make weak learners to learn parallel since random dataset is used for training

whereas Adaboost can make weak learners to learn sequentially since previous misclassified data is given more probability of choosing in the next learning section.

Bagging: A new emerging concept of Ensemble based regression analysis founded on the filtering based ensemble is seen superior to the bootstrap aggregating as studied in (Wei-Liang Tay et al 2012). Bagging method has its own advantage over pruning regression ensembles in which exponential cost is in the size of the ensemble. It is solved using semi definite programming (SDP) or modifying the order of aggregation (Daniel Hernández-Lobato et al 2011). Sub ensembles obtained using either SDP or ordered aggregation usually outperform sub ensembles obtained by other ensemble pruning methods and ensembles generated by the Adaboost.

Adaboost: One of the meta technique, Adaboost Algorithm, is introduced with cost terms into this learning framework (Yanmin Sun et al 2011) leading to the exploration of three models, and one of them tallies with stagewise additive modelling statistics to minimise the cost exponential loss. Thus it adds to an efficient algorithm for resolving imbalanced data problem. Adaboost can incorporate SVM as its component classifier as seen in (Xuchun Li et al 2008), also known as AdaboostSVM outperforms all its counterparts component classifiers such as Decision Trees and Neural Networks. It is under the notion that sequence of trained RBFSVM reduces progressively as the boosting iteration proceeds.

Random Forest: Random Forest has a wide application in which the ensemble classifier can be learned with resampled data (Akin Özçift et al 2011). Since random forest is forest of decision trees, the prediction is enhanced better than a single decision tree. 30 classifier ensembles are constructed based on RF algorithm proved to have accuracy of 87.13% as illustrated in (Akin Ozcift et al 2011). A new extension of random forest known as Dynamic Random Forests (DRF) is studied in (Simon Bernard et al 2012). It is based on a adaptive tree induction procedure such that each tree complement as much as trees possible in RF. It is done through resampling of training data and boosting algorithm and found to produce promising results than the conventional RF. Another new version of RF is the random survival forests (Hemant Ishwaran et al 2010). Consistency of the new method is proved under general splitting rules, bootstrapping

and random selection of variables. It is proved that forest ensemble survival function converges uniformly.

Decorate: Decorate method constructs diverse learners by using artificial data. It works well in cases of missing features, classification noise and feature noise as observed in (Prem Melville et al 2004). Decorate outsmarts Bagging and Adaboost in cases mentioned above. Decorate effectively decreases the error of the base learner.

Combining Methods

Combining methods are employed to evaluate and specify one final result for the ensemble of predictions. Various combining methods of the literature are evaluated in (Lior Rokach et al 2010) and are as follows

Uniform Voting: In the uniform voting, each classifier has the same weight. A classification of an unlabeled instance is performed according to the class that obtains the highest number of votes. Mathematically it can be written as:

$$Class(x) = \underset{c_i \in dom(y)}{\operatorname{argmax}} \sum_{c_j \in dom(y)} 1 \quad \forall k c_i = \underset{c_j \in dom(y)}{\operatorname{argmax}} \hat{P}_{M_k}(y=c_j|x)$$

Where M_k denotes classifier k and $\hat{P}_{M_k}(y=c|x)$ denotes the probability of y obtaining the value c given an instance x .

Distribution Summation: The idea behind distribution summation is to sum up the conditional probability vector obtained from each classifier. The selected class is chosen according to the highest value in the total vector. Mathematically, it can be written as:

$$Class(x) = \underset{c_i \in dom(y)}{\operatorname{argmax}} \sum_k \hat{P}_{M_k}(y=c_i|x)$$

Bayesian Combination: This combining method was investigated by Buntine (1990). The weight associated with each classifier is the posterior probability of the classifier given the training set.

$$Class(x) = \underset{c_i \in dom(y)}{\operatorname{argmax}} \sum_k P(M_k|S) \cdot \hat{P}_{M_k}(y=c_i|x)$$

where $P(M_k|S)$ denotes the probability that the classifier M_k is correct given the training set S . The estimation of $P(M_k|S)$ depends on the classifier's representation.

Dempster-Shafer: The idea of using the Dempster-Shafer theory of evidence (Buchanan and Shortliffe, 1984) for combining models has been suggested by Shilen (1992). This method uses the notion of basic

probability assignment defined for a certain class c_i given the instance x :

$$bpa(c_i, x) = 1 - \prod_k (1 - \hat{P}_{M_k}(y = c_i | x))$$

Subsequently, the selected class is the one that maximizes the value of the belief function:

$$Bel(c_i, x) = \frac{1}{A} \cdot \frac{bpa(c_i, x)}{1 - bpa(c_i, x)}$$

where A is a normalization factor defined as:

$$A = \sum_{\forall c_i \in \text{dom}(y)} \frac{bpa(c_i, x)}{1 - bpa(c_i, x)} + 1$$

Naïve Bayes: Using Bayes' rule, one can extend the Naïve Bayes idea for combining various classifiers:

$$\text{class}(x) = \underset{c_j \in \text{dom}(y)}{\text{argmax}} \quad \hat{P}(y = c_j) \cdot \prod_{k=1} \frac{\hat{P}_{M_k}(y = c_j | x)}{\hat{P}(y = c_j)}$$

$\hat{P}(y = c_j) > 0$

Entropy Weighting: Entropy weighting gives each classifier a weight that is inversely proportional to the entropy of its classification vector.

$$\text{Class}(x) = \underset{c_i \in \text{dom}(y)}{\text{argmax}} \sum_{k: c_i = \underset{c_j \in \text{dom}(y)}{\text{argmax}} \hat{P}_{M_k}(y = c_j | x)} \text{Ent}(M_k, x)$$

where:

$$\text{Ent}(M_k, x) = - \sum_{c_j \in \text{dom}(y)} \hat{P}_{M_k}(y = c_j | x) \log(\hat{P}_{M_k}(y = c_j | x))$$

Logarithmic Opinion Pool: According to the logarithmic opinion pool (Hansen, 2000) the selection of the preferred class is performed according to:

$$\text{Class}(x) = \underset{c_j \in \text{dom}(y)}{\text{argmax}} e^{\sum_k \alpha_k \cdot \log(\hat{P}_{M_k}(y = c_j | x))}$$

where α_k denotes the weight of the k -th classifier, such that:

$$\alpha_k \geq 0; \sum \alpha_k = 1$$

Comparative Study

Random forest performs well in the case of high dimensional data. So enhancement of (Yongqing Zhang et al 2012) can be proposed in which under sampling technique at the data level as well as random forest at algorithmic level can be integrated to benefit a better prediction. Feature selection can be done through auto covariance method, and the base learners can be SVM and ANN as in (Yongqing Zhang et al 2012). The random forest which is a combination of all the decision trees posterior to randomising of datasets. As stated in (Rich Caruana et al 2008), the

following figure 2 suggests the performance of Random forest in high dimensions. Based on the study of (Rich Caruana et al 2008) paves the capability and compatibility of choosing Random forest for (Yongqing Zhang et al 2012) seems an efficient solution over Bagging and Adaboost method

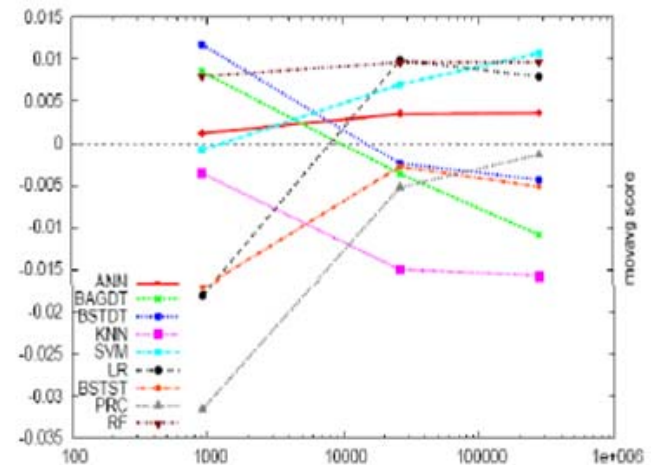


FIG. 2 MOVING AVERAGE SCORE OF EACH LEARNING ALGORITHM AGAINST DIMENSIONS

Conclusion

Numerous solutions to imbalanced data problem is thoroughly studied in this paper. These solutions have been classified under various level such as data and algorithmic level. A detailed study of one paper led to the conclusion that there is a scope for modifying Bagging and Adaboost with Random Forest method as it can deal with high dimensional data very well based on the extensive study made on this domain. As a future work comparative evaluation of ensemble of ensemble classifiers with high dimensional data can be studied.

REFERENCES

- Akin Özçift, May 2011, Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis, Computers in Biology and Medicine, Volume 41, Issue 5, Pages 265-271.
- Akin Ozcift, Arif Gulen, December 2011, Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms, Computer Methods and Programs in Biomedicine, Volume 104, Issue 3, Pages 443-451.
- Alberto Fernández, María José del Jesus, Francisco Herrera, August 2009, On the influence of an adaptive inference system in fuzzy rule based classification systems for

- imbalanced data-sets, *Expert Systems with Applications*, Volume 36, Issue 6, Pages 9805-9812.
- Asifullah Khan, Abdul Majid, Maqsood Hayat, August 2011, CE-PLoc: An ensemble classifier for predicting protein subcellular locations by fusing different modes of pseudo amino acid composition, *Computational Biology and Chemistry*, Volume 35, Issue 4, Pages 218-229.
- Chan-Yun Yang, Jr-Syu Yang, Jian-Jun Wang, December 2009, Margin calibration in SVM class-imbalanced learning, *Neuro computing*, Volume 73, Issues 1-3, Pages 397-411.
- Charles Elkan, 2001, The Foundations of Cost-Sensitive Learning, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*.
- Chien-I Lee, Cheng-Jung Tsai, Tong-Qin Wu, Wei-Pang Yang, May 2008, An approach to mining the multi-relational imbalanced database, *Expert Systems with Applications*, Volume 34, Issue 4, Pages 3021-3032.
- Chou-Yuan Lee, Zne-Jung Lee, August 2012, A novel algorithm applied to classify unbalanced data, *Applied Soft Computing*, Volume 12, Issue 8, Pages 2481-2485.
- Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz, Alberto Suárez, June 2011, Empirical analysis and evaluation of approximate techniques for pruning regression bagging ensembles, *Neurocomputing*, Volume 74, Issues 12-13, Pages 2250-2264.
- Der-Chiang Li, Chiao-Wen Liu, Susan C. Hu, May 2010, A learning method for the class imbalance problem with medical data sets, *Computers in Biology and Medicine*, Volume 40, Issue 5, Pages 509-518.
- Erika Antal, Yves Tillé, January 2011, Simple random sampling with over-replacement, *Journal of Statistical Planning and Inference*, Volume 141, Issue 1, Pages 597-601.
- Francisco Fernández-Navarro, César Hervás-Martínez, Pedro Antonio Gutiérrez, August 2011, A dynamic over-sampling procedure based on sensitivity for multi-class problems, *Pattern Recognition*, Volume 44, Issue 8, Pages 1821-1833.
- Godfried T Toussaint, August 1994, A counterexample to Tomek's consistency theorem for a condensed nearest neighbor decision rule, *Pattern Recognition Letters*, Volume 15, Issue 8, Pages 797-801.
- Hemant Ishwaran, Udaya B. Kogalur, July 2010, Consistency of random survival forests, *Statistics & Probability Letters*, Volume 80, Issues 13-14, 1-15, Pages 1056-1064.
- Hung-Yi Lin, June 2012, Efficient classifiers for multi-class classification problems, *Decision Support Systems*, Volume 53, Issue 3, Pages 473-481.
- Jae Pil Hwang, Seongkeun Park, Euntai Kim, July 2011, A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function, *Expert Systems with Applications*, Volume 38, Issue 7, Pages 8580-8585.
- Jin Xiao, Ling Xie, Changzheng He, Xiaoyi Jiang, 15 February 2012, Dynamic classifier ensemble model for customer classification with imbalanced class distribution, *Expert Systems with Applications*, Volume 39, Issue 3, Pages 3668-3675.
- Jinghua Wang, Jane You, Qin Li, Yong Xu, March 2012, Extract minimum positive and maximum negative features for imbalanced binary classification, *Pattern Recognition*, Volume 45, Issue 3, Pages 1136-1145.
- Larry Shoemaker, Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, W. Philip Kegelmeyer, January 2008, Using classifier ensembles to label spatially disjoint data, *Information Fusion*, Volume 9, Issue 1, Pages 120-133.
- Lior Rokach, 2010, Ensemble methods for classifiers, *Data Mining and Knowledge Discovery Handbook*.
- M.A.H. Farquad, Indranil Bose Preprocessing unbalanced data using support vector machine, *Decision Support Systems*, Volume 53, Issue 1, April 2012, Pages 226-233.
- María Dolores Pérez-Godoy, Alberto Fernández, Antonio Jesús Rivera, María José del Jesus, November 2010, Analysis of an evolutionary RBFN design algorithm, CO²RBFN, for imbalanced data sets, *Pattern Recognition Letters*, Volume 31, Issue 15, Pages 2375-2388.
- Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, Francisco Herrera, August 2011, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, *Pattern Recognition*, Volume 44, Issue 8, Pages 1761-1776.

- Ming Gao, Xia Hong, Sheng Chen, Chris J. Harris, October 2011, A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems, *Neurocomputing*, Volume 74, Issue 17, Pages 3456-3466.
- Mu-Chen Chen, Long-Sheng Chen, Chun-Chin Hsu, Wei-Rong Zeng, August 2008, An information granulation based data mining approach for classifying imbalanced data, *Information Sciences*, Volume 178, Issue 16, Pages 3214-3227.
- Muhammad Atif Tahir, Josef Kittler, Fei Yan, October 2012, Inverse random under sampling for class imbalance problem and its application to multi-label classification, *Pattern Recognition*, Volume 45, Issue 10, Pages 3738-3750.
- Nicolás García-Pedrajas, César García-Osorio, January 2011, Constructing ensembles of classifiers using supervised projection methods based on misclassified instances, *Expert Systems with Applications*, Volume 38, Issue 1, Pages 343-359.
- Noritaka Shigei, Hiromi Miyajima, Michiharu Maeda, Lixin Ma, December 2009, Bagging and AdaBoost algorithms for vector quantization, *Neurocomputing*, Volume 73, Issues 1-3, Pages 106-114.
- Paolo Soda, August 2011, A multi-objective optimisation approach for class imbalance learning, *Pattern Recognition*, Volume 44, Issue 8, Pages 1801-1810.
- Pilsung Kang, Sungzoon Cho, Douglas L. MacLachlan, June 2012, Improved response modeling based on clustering, under-sampling, and ensemble, *Expert Systems with Applications*, Volume 39, Issue 8, Pages 6738-6753.
- Pooja Jain, Jonathan M. Garibaldi, Jonathan D. Hirst, June 2009, Supervised machine learning algorithms for protein structure classification, *Computational Biology and Chemistry*, Volume 33, Issue 3, Pages 216-223.
- Prem Melville, Nishit Shah, Lilyana Mihalkova, Raymond J. Mooney, June 2004, Experiments on Ensembles with Missing and Noisy Data, *Proceedings of 5th International Workshop on Multiple Classifier Systems (MCS-2004)*, LNCS Vol. 3077, pp. 293-302, Cagliari, Italy, Springer Verlag.
- Raskutti, B., Kowalczyk, A., 2004. Extreme re-balancing for SVMs: a case study. *ACM SIGKDD Explorations Newsletter* 6, 60-69.
- Rich Caruana, Nikos Karampatziakis, Ainur Yessenalina, 2008, An Empirical Evaluation of Supervised Learning in High Dimensions, *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.
- Rongsheng Gong, Samuel H. Huang, May 2012, A Kolmogorov-Smirnov statistic based segmentation approach to learning from imbalanced datasets: With application in property refinance prediction, *Expert Systems with Applications*, Volume 39, Issue 6, Pages 6192-6200.
- Salvador García, Joaquín Derrac, Isaac Triguero, Cristóbal J. Carmona, Francisco Herrera, February 2012, Evolutionary-based selection of generalized instances for imbalanced classification, *Knowledge-Based Systems*, Volume 25, Issue 1, Pages 3-12.
- Show-Jane Yen, Yue-Shi Lee, April 2009, Cluster-based under-sampling approaches for imbalanced data distributions, *Expert Systems with Applications*, Volume 36, Issue 3, Part 1, Pages 5718-5727.
- Show-Jane Yen, Yue-Shi Lee, April 2009, Cluster-based under-sampling approaches for imbalanced data distributions, *Expert Systems with Applications*, Volume 36, Issue 3, Part 1, Pages 5718-5727.
- Shu-xue Zou, Yan-xin Huang, Yan Wang, Chun-guang Zhou, September 2008, A Novel Method for Prediction of Protein Domain Using Distance-Based Maximal Entropy, *Journal of Bionic Engineering*, Volume 5, Issue 3, Pages 215-223.
- Simon Bernard, Sébastien Adam, Laurent Heutte, September 2012, Dynamic Random Forests, *Pattern Recognition Letters*, Volume 33, Issue 12, Pages 1580-1586.
- Todd W. Schiller, Yixin Chen, Issam El Naqa, Joseph O. Deasy, June 2010, Modeling radiation-induced lung injury risk with an ensemble of support vector machines, *Neurocomputing*, Volume 73, Issues 10-12, Pages 1861-1867.
- Victoria López, Alberto Fernández, Jose G. Moreno-Torres, Francisco Herrera, June 2012, Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics, *Expert Systems with Applications*, Volume 39, Issue 7, Pages 6585-6608.

- Wei-Liang Tay, Chee-Kong Chui, Sim-Heng Ong, Alvin Choong-Meng Ng, August 2012, Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis, *Expert Systems with Applications*.
- Xuchun Li, Lei Wang, Eric Sung, August 2008, AdaBoost with SVM-based component classifiers, *Engineering Applications of Artificial Intelligence*, Volume 21, Issue 5, Pages 785-795.
- Yang Liu, Xiaohui Yu, Jimmy Xiangji Huang, Aijun An, July 2011, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets, *Information Processing & Management*, Volume 47, Issue 4, Pages 617-631.
- Yang Yong, 2012, The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm, *Energy Procedia*, Volume 17, Part A, Pages 164-170.
- Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, Yang Wang, December 2011, Cost-sensitive boosting for classification of imbalanced data, *Pattern Recognition*, Volume 40, Issue 12, Pages 3358-3378.
- Yok-Yen Nguwi, Siu-Yeung Cho, "An unsupervised self-organizing learning with support vector ranking for imbalanced datasets", *Expert Systems with Applications*, Volume 37, Issue 12, Pages 8303-8312, December 2010.
- Yongqing Zhang, Danling Zhang, Gang Mi, Daichuan Ma, Gongbing Li, Yanzhi Guo, Menglong Li, Min Zhu, February 2012, Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions, *Computational Biology and Chemistry*, Volume 36, Pages 36-41.
- Young-Won Kim, Il-Seok Oh, April 2008, Classifier ensemble selection using hybrid genetic algorithms, *Pattern Recognition Letters*, Volume 29, Issue 6, Pages 796-802.
- Zhuangyuan Zhao, Ping Zhong, Yaohong Zhao, August 2011, Learning SVM with weighted maximum margin criterion for classification of imbalanced data, *Mathematical and Computer Modelling*, Volume 54, Issues 3-4, Pages 1093-1099.